UNIFIED MEASURES FOR THE RATE-DISTORTION-LATENCY TRADE-OFF

Melan Vijayaratnam[1] *Marta Milovanović*[1] *Marco Cagnazzo*[1,2]
Enzo Tartaglione[1] *Giuseppe Valenzise*[3]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]University of Padua, Department of Information Engineering, Italy
[3]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France

ABSTRACT

In today's digital age, multimedia content is omnipresent, and the demand for efficient compression techniques is ever-increasing. In particular, the successful delivery of services based on video transmission largely depends on achieving the lowest latency values. One solution has been to use extrapolation for latency compensation in video transmission that allows to reduce the latency by an arbitrary amount. Nevertheless, this latency reduction comes at the cost of an increased distortion of the displayed images, since they are based on temporal extrapolation. Latency can also be traded with coding rate. This paper introduces ELR-PSNR and EPR-Latency as unified metrics to assess the three-way trade-off between rate, distortion, and latency simultaneously.

***Index Terms*—** low-latency video delivery, metric, rate-distortion-latency trade-off

1. INTRODUCTION

Achieving ultra low-latency video delivery is a crucial requirement in numerous applications involving human interactions (*e.g.*, video conferencing, virtual and augmented reality) or human-machine interactions (*e.g.*, teleoperation of unmanned vehicles or robots, *etc*.). In these scenarios, the Glass-to-Glass latency, which represents the delay between the acquisition of a video frame by one agent and its display by a second remote agent, holds significant importance [1]. This latency factor heavily influences the overall quality of experience perceived by users [2].

The latency compensation framework [3], approaches the problem of ultra-low-latency video transmission between a transmitter and a receiver. To mitigate the impact of G2G latency, a compensatory approach involves extrapolating the available present information on the receiver side to predict and display future frames prior to their actual reception, displayed in Figure 1. To assess performance, the Bjøntegaard metric [4] is used to compute the Bjøntegaard delta (BD)-PSNR which measures the change in the peak signal-to-noise ratio or PSNR [5] to achieve a certain level of rate improvement compared to a reference codec. The structure similarity index (SSIM) [6] or the recent video multi-method assessment fusion (VMAF) [7] can be similary used for BD-SSIM and BD-VMAF calculations to account for the different distortion measures.

While the BD-PSNR, BD-SSIM, BD-VMAF are effective metrics for analyzing rate-distortion trade-offs, they do not explicitly account for latency considerations. This paper proposes a novel three-way metric for the rate-distortion-latency trade-off. Our metrics expand upon traditional rate-distortion analysis by including latency as a crucial factor in assessing compression algorithms. We introduce two new metrics:

- ELR-PSNR: Equal-Latency, Equal-Rate Delta PSNR
- EPR-Latency: Equal-PSNR, Equal-Rate Delta Latency

which aim to characterize the three-way trade-off among rate, distortion, and latency (RDL).

2. RELATED WORK

Introduced in 2001, the Bjøntegaard Delta (BD) method continues to be one of the most commonly utilized tools for calculating and comparing the compression efficiency of video codecs. It has been used to evaluate and compare thoroughly the most recent codecs [8]. Alexis *et al.* [9] presented an Excel template for calculated BD bitrate values with more than four data points and additional modes to handle cases where there are overlapping issues. Later, Herglotz *et al.* [10] proposed the Akima interpolation for more accurate results.

As the pursuit of more efficient codec designs advances, it naturally leads to the investigation of calculating coding efficiency across various codec options. The SCENIC metric [11] relies on the mean opinion score as opposed as the PSNR. Later, the Bjøntegaard-Delta decoding energy (BDDE) [12] is introduced to describe the energy savings in % for the same PSNR. On the other hand, we also take into account latency to propose unified measures for the rate-distortion-latency trade-off.

Video extrapolation is a technique used to predict and generate future video frames based on the available information in the existing video sequence. There are various approaches

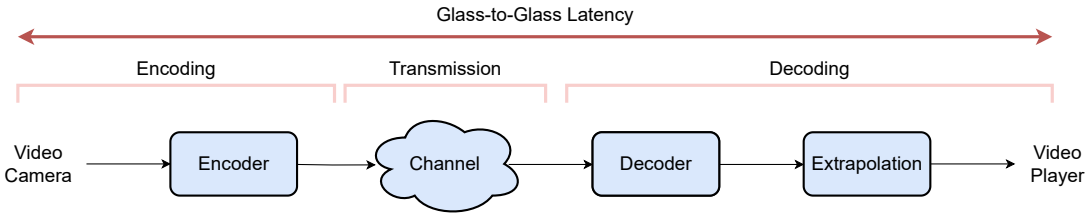

Fig. 1: Latency compensation scheme when extrapolation is performed at the decoder side to reduce the Glass-to-Glass latency.

to video extrapolation, including pixel-based, motion-based and fusion-based methods as pinpointed by Gao *et al.* [13]. All of these extrapolation methods are an essential part for the latency compensation framework [3].

3. METHODOLOGY

3.1. Definitions and notations

The ELR-PSNR and EPR-Latency metrics offer a versatile approach to assess the trade-off among latency, rate, and distortion for various coding methods, extending their utility beyond cases involving latency control through extrapolation. This includes applications such as comparing configurations of different codecs. For instance, the metrics can be used to compare Random Access (RA) with Low Delay P (LDP). In fact, in order to evaluate the RDL trade-off between two methods, be them method A and method B, we only need to collect, for both of them, a suitable number of operation points, which is a triplet of rate R, distortion D and latency L values associated to one specific execution of the method. In order to simplify the description, let us first consider the case where the video encoder does not change. We could consider for example HEVC in Low Delay P (LDP) configuration. We run this baseline with N^Q values of QP and store the corresponding RD points in two vectors:

$$\begin{aligned} R &= [R(1), R(2), \ldots, R(N^Q)]^T \\ D &= [D(1), D(2), \ldots, D(N^Q)]^T \end{aligned} \tag{1}$$

Let us say we run method A over an input video sequence. We consider N_A^Q values of the quantization parameter QP and N_A^h values of the extrapolation horizon h. If the considered method A does not allow frame prediction, we consider $N_A^h = 1$ and $h = 0$ meaning no prediction in the future is performed. If the method allows frame prediction typical values of N_h include 5 to 10. Thus, for each method we collect $N_A = N_A^Q \cdot N_A^h$ operation points. We arrange them as follows:

$$\begin{aligned} R_A &= [R_A(Q_k, h_p)]_{\substack{1 \le k \le N_A^Q \\ 1 \le p \le N_A^h}} \\ D_A &= [D_A(Q_k, h_p)]_{\substack{1 \le k \le N_A^Q \\ 1 \le p \le N_A^h}} \\ L_A &= [L_A(Q_k, h_p)]_{\substack{1 \le k \le N_A^Q, \\ 1 \le p \le N_A^h}} \end{aligned} \tag{2}$$

where k represents the values for the different quantization parameters ordered in a column, and p the values for the different extrapolation horizons. The latency should be computed as a difference with respect to the baseline. For a given Q_k, we get the following equation:

$$L_A^{Q_k}(h_i) = T_E^{A,Q_k}(h_i) - T_F \cdot h_i \tag{3}$$

where h_i is the temporal extrapolation horizon, $T_E^{A,Q_k}(n)$ is the time needed by method A to compute an extrapolated frame with horizon n, and $T_F = 1/F$ is the frame interval, inverse of the frame rate F. The extrapolation horizon h_i expresses the number of frames predicted in the future and therefore translates directly to the latency compensated based on the individual latency that each method induces. This idea can be generalized to measure any kind or any part of the latency, such as the encoding time for different codec presets in video codecs. Likewise, we gather $N_B = N_B^Q \cdot N_B^h$ operation points for methods B and we arrange R_B, D_B and L_B analogously as explained for the method A in Equation (2).

3.2. ELR-PSNR and EPR-Latency

The quality saving difference between the two rate-distortion-latency surfaces at a given level of fidelity is :

$$\Delta D(R, L) = \frac{D_B(R, L) - D_A(R, L)}{D_A(R, L)} \tag{4}$$

where $D_A(R, L)$ and $D_B(R, L)$ are respectively the distortion of the interpolated of the methods A and B at a given level of rate R and Latency L. $\Delta D(R, L)$ is positive for a gain in quality whereas a negative value indicates a degradation in quality of method A with respect to method B.

Similar to the Bjøntegaard model, we use a logarithmic scale for the domain of the quality interpolation, so by defining $d = \log D$ the ELR-PSNR savings can be expressed as:

$$\Delta D(R, L) = 10^{d_B(R,L) - d_A(R,L)} - 1 \tag{5}$$

By considering the measured rate-distortion-latency points $(R(i, j), D(i, j), L(i, j))$, the computation of the ELR-PSNR involves utilizing the fitted rate-distortion-latency surfaces $\hat{d}(R, L)$. The ELR-PSNR approximation is then determined over a specified range of rate levels, determined by the rate,

and latencies:

$$\Delta D_{Overall} \approx 10^{\frac{1}{(R_h-R_l)\cdot(L_h-L_l)}\int_{R_l}^{R_h}\int_{L_l}^{L_h}[\hat{d}_B(R,L)-\hat{d}_A(R,L)]dRdL} - 1 \quad (6)$$

The lower integration bounds R_l and L_l, and higher integration bounds R_h and L_h are derived from the range of interpolated rate and latency values from method A and B:

$$\begin{aligned} R_h &= max\{min(R_A, R_B)\} \\ R_l &= min\{max(R_A, R_B)\} \\ L_h &= max\{min(L_A, L_B)\} \\ L_l &= min\{max(L_A, L_B)\} \end{aligned} \quad (7)$$

Likewise to have a measure about the latency gain between two rate-distortion-latency surfaces at a given level of latency, we introduce the EPR-Latency computed as follows:

$$\Delta L_{Overall} \approx 10^{\frac{1}{(R_h-R_l)\cdot(D_h-D_l)}\int_{R_l}^{R_h}\int_{D_l}^{D_h}[\hat{l}_B(R,D)-\hat{l}_A(R,D)]dRdD} - 1 \quad (8)$$

with D_l and D_h respectively the lower and higher integration bounds of distortion of methods A and B and $\hat{l}(R, D)$ the fitted rate-distortion surfaces from the measured rate-distortion-latency points.

4. EXPERIMENTAL RESULTS

4.1. Test conditions

For the video extrapolation methods, we select MCNet [14] as the pixel-based method, FlowNet2 [15] as a the motion-based method and SDCNet [16] as the fusion-based methods. FlowNet2 is the sole supervised approach, which means it requires optical flow labels and can only use pretrained weights, in contrast to other methods that can be retrained on any dataset because of their self-supervised categorization. "SDCNet iter" refers to the basic SDCNet architecture that iteratively re-circulate the last predicted output back as input and "SDCNet direct" uses temporal subsampling to predict directly the frame h steps in the future. The temporal subsampling provides a better rate-distortion-latency trade-off we will show later. FlowNet2 also uses recirculation to predict beyond $h > 1$ and employs a warping operation that applies computed flows to previous frame pixels for next-frame prediction. In contrast, MCNet is based on long short-term memory and makes multiple simultaneous predictions.

For training, we use the UCF101 action recognition dataset of realistic action videos, collected from Youtube [17]. The training set consists in 127,654 images from sequences containing sufficient temporal information to allow the neural network to learn the motion. The test set contain 10 sequences cut at 250 frames of spanning over different set of actions.

	BD-PSNR ↑		
horizon h	**1**	**3**	**5**
Copylast	-13.19	-17.43	-18.84
MCNet	-11.37	-16.78	-18.79
FlowNet2 + warp	-12.17	-16.96	-18.69
SDCNet iter	-9.69	-14.03	-15.82
SDCNet direct	-8.67	-15.47	-17.32

(a) HEVC All Intra results

	BD-PSNR ↑		
horizon h	**1**	**3**	**5**
Copylast	-10.72	-14.79	-16.18
MCNet	-8.92	-12.29	-16.07
FlowNet2 + warp	-9.72	-14.30	-15.99
SDCNet iter	-7.37	-11.38	-13.12
SDCNet direct	-7.37	-12.82	-14.65

(b) HEVC LDP results

	BD-PSNR ↑		
horizon h	**1**	**3**	**5**
Copylast	-11.13	-15.37	-16.80
MCNet	-9.32	-14.74	-16.75
FlowNet2 + warp	-10.08	-14.86	-16.62
SDCNet iter	-7.51	-11.69	-13.52
SDCNet direct	-7.51	-13.22	-15.17

(c) VVenC RA results

Table 1: Quantitative results on UCF101 sequences

All sequences are captured at 25 frame per second with the resolution 256×256.

To evaluate the efficacy on the latency compensation scheme presented in Figure 1, we employ the HEVC and VVC codecs as evaluation benchmarks. In the case of the HEVC codec, we utilize both the All-Intra (AI) and Low Delay P (LDP) configurations, employing the HEVC HM codec implementation as documented in [18]. As for the VVC codec, we focus on the Random Access (RA) configuration, utilizing the VVenC implementation as described in [19].

4.2. Beyond Bjøntegaard

To evaluate the latency compensation scheme [3], the different codecs and configurations are used with the quantization parameters $QP \in \{22, 27, 32, 37\}$ for all the selected methods. Rate-Distortion (RD) points are computed considering an extrapolation horizon $h \in \{1, 2, 3, 4, 5\}$. All RD curves are compared to the case $h = 0$ which means that no extrapolation is performed. Table 1 reports the BD-PSNR results of the different methods across the different horizon h displayed at horizon 1, 3, 5 for conciseness.

The issue with such an approach is that we consider the rate-distortion trade-off at separate horizon h resulting in many measurement necessary. Furthermore, the use of the extrapolation horizon h is necessary to compare the different numbers, if we were to use latency values, predicting at horizon $h = 3$ steps in the future corresponds to a latency compensated of $L = 3 \times 16.7 - 120 = -69.9$ ms for "SD-

	HEVC All Intra		HEVC LDP		VVenC RA	
	ELR-PSNR [dB] ↑	**EPR-Latency [ms] ↑**	**ELR-PSNR [dB] ↑**	**EPR-Latency [ms] ↑**	**ELR-PSNR [dB] ↑**	**EPR-Latency [ms] ↑**
MCNet	-2.27	-64.37	-2.13	-61.84	-2.30	-64.00
Flownet2 + warp	-1.50	-30.24	-1.33	-27.93	-1.39	-27.78
SDCNet iter	-2.22	-24.59	-2.25	-22.00	-2.27	-21.72
SDCNet direct	0.90	20.97	0.55	22.86	0.63	24.19

Table 2: ELR-PSNR and EPR-Latency for UCF101 sequences.

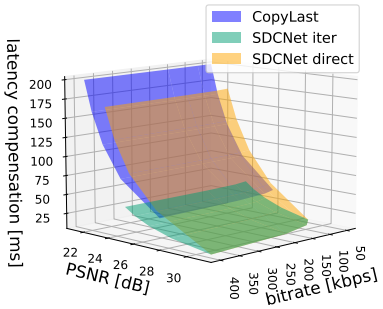

Fig. 2: Rate-distortion-latency surfaces for Copylast, "SDCNet iter" and "SDCNet direct". We see the impact iterative method of SDCNet and its impact on the compensated latency despite better quality than the direct method.

CNet iter" whereas it corresponds to a latency compensation using Equation 3 of $L = 16.7 - 120 = -103.3$ ms for "SDCNet direct". These latencies emphasize the distortion-latency trade-off of what the the temporal subsampling applied to SDCNet entails: a loss in quality but resulting in more latency compensation. Here we can clearly see the limitations of considering only rate-distortion curves.

4.3. The rate-distortion-latency trade-off

We display the 3-dimensional surfaces for CopyLast, "SDCNet iter" and "SDCNet direct" in Figure 2. We do so by interpolating the points of the matrices R_A, D_A, L_A in equations 2 from method A and R_B, D_B, L_B from method B with the fitted rate-distortion-latency surfaces $\hat{d}_A(R, L)$ and $\hat{d}_B(R, L)$. The latency compensation takes into account the extrapolation time measured by a NVIDIA Geforce RTX 3090. For a set of fixed PSNR and bitrate points, the surface is calculated between method A and B giving thus the EPR-Latency value.

By switching to a three-dimensional perspective it is therefore possible to compare rate-distortion curves at multiple horizon h at once. In Table 2 we show the ELR-PSNR and EPR-Latency of the different extrapolation methods with CopyLast. We see clearly that "SDCNet direct" provides true quality gain of 0.55 dB and 22.86 ms of latency gain using HEVC LDP configuration whereas the other methods report loss. This is due to the fact that the extrapolation time is negatively impacting the ELR-PSNR and EPR-Latency despite

HEVC LDB -	**ELR-PSNR [dB] ↑**	**EPR-Latency [s] ↑**
HEVC LDP	0.03	153.37
HEVC RA	1.03	261.69
VVenC slow	2.01	150.00
VVenC slower	0.92	144.85

Table 3: Comparing codecs presets with the ELR-PSNR and EPR-Latency by taking into account the encoding time as latency.

the fact that Table 1 suggests that there will be less loss while using "SDCNet iter". However, because the extrapolation time is linear with the extrapolation horizon h when using "SDCNet iter", the quality gain compared to "SDCNet direct" does not compare to its constant extrapolation time with respect to h.

In Table 3 we show that the ELR-PSNR and EPR-Latency can be applied to a different context. Here we compare the latency-rate-distortion trade-of between two different presets, with HEVC Low Delay B (LDB) with different configurations. The latency here is represented by the encoding time measured and the table shows how the other methods are faster in encoding compared to HEVC LDB.

5. CONCLUSION

In this paper we proposed two unified measures for the rate-distortion-latency trade-off.The ELR-PSNR and EPR-Latency for the rate-distortion-latency trade-off represent, a development in the multimedia compression industry. By concurrently taking into account the interplay between rate, distortion, and latency, which are critical elements in assessing the efficacy of compression algorithms, this innovative metric tackles the shortcomings of conventional techniques.

Future research can refine and validate the three-way metric across different compression algorithms and applications. Additionally, investigating the potential trade-offs and optimizing strategies within the three-way trade-off space would contribute to further advancements in multimedia compression metrics and techniques.

6. ACKNOWLEDGMENTS

This work was funded by the ANR AAPG2020 national fund (ANR-20-CE25-0014). I would like to thank Nikolai Fadeev for his helpful suggestions.

7. REFERENCES

[1] C. Bachhuber, E. Steinbach, M. Freundl, and M. Reisslein, "On the Minimization of Glass-to-Glass and Glass-to-Algorithm Delay in Video Communication," *IEEE Trans. on Multimedia*, vol. 20, no. 1, 2018.

[2] K. Brunnström, E. Dima, T. Qureshi, M. Johanson, M. Andersson, and M. Sjöström, "Latency impact on quality of experience in a virtual reality simulator for remote control of machines," *Signal Processing: Image Communication*, vol. 89, pp. 116005, 2020.

[3] M. Vijayaratnam, M. Cagnazzo, G. Valenzise, A. Trioux, and M. Kieffer, "Towards zero-latency video transmission through frame extrapolation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022.

[4] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

[5] G. Sullivan and K. Minoo, "Objective quality metric and alternative methods for measuring coding efficiency," in *document JCTVC-H0012, ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), 8th Meeting: San Jose, CA, USA*, 2012.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, Apr. 2004.

[7] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, Aug. 2019.

[8] T. K. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J.-R. Ohm, and G. J. Sullivan, "Video quality evaluation methodology and verification testing of hevc compression performance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, 2015.

[9] Y. Ye, E. Alshina, and J. Boyce, "Joint video exploration team (jvet) of itu-t sg 16 wp 3 and iso/iec jtc 1/sc 29/wg 11," .

[10] C. Herglotz, M. Kränzler, R. Mons, and A. Kaup, "Beyond bjøntegaard: Limits of video compression performance comparisons," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 46–50.

[11] P. Hanhart and T. Ebrahimi, "Calculation of average coding efficiency based on subjective quality scores," *Journal of Visual communication and image representation*, vol. 25, no. 3, pp. 555–564, 2014.

[12] M. Kränzler, C. Herglotz, and A. Kaup, "Energy efficient video decoding for vvc using a greedy strategy-based design space exploration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, 2021.

[13] H. Gao, H. Xu, Q.-Z. Cai, R. Wang, F. Yu, and T. Darrell, "Disentangling propagation and generation for video prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[14] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[15] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[16] F. A. Reda, G. Liu, K. J. Shih, R. Kirby, J. Barker, D. Tarjan, A. Tao, and B. Catanzaro, "Sdc-net: Video prediction using spatially-displaced convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[17] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[18] C. Rosewarne, K. Sharman, R. Sjöberg, and G. J. Sullivan, "High Efficiency Video Coding (HEVC) Test Model 16 (HM 16) Encoder Description Update 13 | MPEG," in *38th JCT-VC Meeting*, Brussels, Jan. 2020.

[19] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, "Vvenc: An open and optimized vvc encoder implementation," in *Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2021.